PENGEMBANGAN APLIKASI ANALISIS SENTIMEN *TWITTER* MENGGUNAKAN METODE *NAÏVE BAYES CLASSIFIER* (Studi Kasus SAMSAT Kota Malang)

Imam Fahrur Rozi[1], Elok Nur Hamdana[2], Muhammad Balya Iqbal Alfahmi[3]

[1,2]Program Studi Teknik Informatika, Teknologi Informasi, [3]Politeknik Negeri Malang
[1]imam.rozi@gmail.com, [2]elokhamdana@gmail.com, [3]iqbalalfahmii@gmail.com

Abstrak

Twitter adalah salah satu media sosial dimana pengguna dapat mencari topik tertentu dan membahas isu-isu terkini. Beberapa pesan singkat atau *tweet* dapat memuat opini terhadap produk dan layanan yang dirasakan oleh masyarakat. Data ini dapat menjadi sumber data untuk dijadikan objek penelitian. Penelitian ini bertujuan untuk membangun aplikasi analisis sentimen yang menerapkan pendekatan *Naïve Bayes Classifier* untuk mengklasifikasikan kata-kata dan difokuskan pada *tweet* dalam bahasa Indonesia. Data diperoleh melalui cara *web scrapping* dan sumber teks yang digunakan sebagai topik bahasan adalah Sistem Administrasi Manunggal Satu Atap (SAMSAT) Malang Kota. Proses klasifikasi dilakukan melalui serangkaian tahapan seperti preproses (*case folding*, *cleaning*, *tokenizing*, dan *stopword*) serta proses klasifikasi dengan algoritma *Naïve Bayes Classifier* itu sendiri untuk mendapatkan hasil klasifikasi dengan kategori positif, negatif atau netral. Berdasarkan hasil penelitian, algoritma *Naïve Bayes Classifier* memberikan unjuk kerja yang baik dalam analisis sentimen. Dari hasil uji akurasi klasifikasi yang dilakukan oleh aplikasi menghasilkan nilai akurasi tertinggi pada setiap kategori positif, negatif, netral masing-masing sebesar 82%, 92%, 80% dengan jumlah data latih 200 *tweet* negatif, 200 *tweet* positif, dan 200 *tweet* netral.

Kata kunci : *Text Mining*, *Web Scrapping, Prepocessing, Naïve Bayes Classifier*

1. Pendahuluan

Perkembangan teknologi informasi yang semakin meningkat memberi dampak pada pertukaran informasi dan komunikasi yang semakin mudah. Hal ini ditandai dengan dengan munculnya media sosial seperti *Twitter, Facebook, Yahoo, Google, Youtube, Instagram, Path*. Pertumbuhan media sosial ini juga mendorong adanya informasi tekstual yang besar sehingga muncul kebutuhan penyajian data yang memudahkan pengguna mendapatkan informasi yang akurat. Media sosial *twitter* merupakan salah satu media komunikasi populer saat ini. Hal ini terlihat dari peningkatan pengguna *twitter* yang tercatat di seluruh dunia.

Pertumbuhan *twitter* terus meningkat setiap waktu, sehingga hal tersebut dimanfaatkan para pengguna *twitter* untuk menyampaikan informasi berupa kritik maupun saran kepada Sistem Administrasi Manuggal Satu Atap (SAMSAT) Malang Kota dengan lebih mudah. Semakin banyak pendapat atau keluhan dari masyarakat dapat membentuk opini masyarakat, dan dapat dijadikan masukan terhadap penilaian kinerja layanan Sistem Administrasi Manuggal Satu Atap (SAMSAT) Malang Kota. Analisis sentimen dapat membantu untuk memperoleh gambaran umum presepsi masyarakat dengan mengelompokkan jenis opnin menjadi kategori positif, negatif, atau netral.

Pendekatan yang digunakan dalam penelitian ini merupakan pendekatan yang mengacu pada teorema *Bayes* yang menggunakan prinsip peluang statistika untuk mengkombinasikan pengetahuan sebelumnya dengan pengetahuan baru. Prinsip ini kemudian digunakan untuk memecahkan masalah klasifikasi. Sistem yang dikembangkan dapat melakukan pengklasifikasian data *tweet* yang mengandung kata sentimen yang bersifat positif, negatif, maupun netral. Setelah data terklasifikasikan, kemudian akan dilakukan perhitungan prosentase akurasi hasil klasifikasi sentimen *tweet* terhadap Sistem Administrasi Manuggal Satu Atap (SAMSAT) Malang Kota.

2. Tinjauan Pustaka

2.1. Sentimen analisis

Tugas dasar dalam analisis sentimen adalah mengelompokkan polaritas dari teks yang ada dalam dokumen, kalimat, atau fitur/tingkat aspek dan menentukan apakah pendapat yang dikemukakan dalam dokumen, kalimat atau fitur entitas/aspek bersifat positif, negatif atau netral.

Lebih lanjut *sentiment analysis* dapat menyatakan emosional sedih, gembira, atau marah Liu, B (2012).

2.2. *Web Scrapping*

Web Scraping Turland (2010) adalah proses pengambilan sebuah dokumen semi-terstruktur dari internet, umumnya berupa halaman-halaman *web* dalam bahasa markup seperti *HTML* atau *XHTML*, dan menganalisis dokumen tersebut untuk diambil data tertentu dari halaman tersebut untuk digunakan bagi kepentingan lain. *Web scraping* memiliki sejumlah langkah, sebagai berikut Josi, A. (2014):

a. *Create Scraping Template*: Pembuat program mempelajari dokumen *HTML* dari *website* yang akan diambil informasinya untuk tag *HTML* yang mengapit informasi yang akan diambil.
b. *Explore Site Navigation*: Pembuat program mempelajari teknik navigasi pada *website* yang akan diambil informasinya untuk ditirukan pada aplikasi *web scraper* yang akan dibuat.
c. *Automate Navigation and Extraction*: Berdasarkan informasi yang didapat pada langkat 1 dan 2 di atas, aplikasi *web scraper* dibuat untuk mengotomatisasi pengambilan informasi dari *website* yang ditentukan.
d. *Extracted Data and Package History*: Informasi yang didapat dari langkah 3 disimpan dalam tabel atau tabel-tabel *database*.

2.3. *Naïve Bayes Classifier*

Naive Bayes merupakan teknik prediksi berbasis probabilistik sederhana yang berdasar pada penerapan teorema *Bayes* (aturan *Bayes*) dengan asumsi *independensi* (tidak ketergantungan) yang kuat (naif). Dengan kata lain, dalam *Naive Bayes* model yang digunakan adalah "model fitur *independen*" Prasetyo (2012).

Naïve Bayes Classifer adalah salah satu algoritma yang digunakan untuk klasifikasi teks serta merupakan metode *Machine Learning* yang menggunakan perhitungan probabilitas dan statistik yang dikemukakan oleh *Thomas Bayes*. Algoritma tersebut digunakan untuk memprediksi probabilitas di masa depan berdasarkan pengalaman di masa lalu.

Dasar dari *naïve bayes* yang dipakai adalah rumus :

$$P(A|B) = \frac{P(B|A) * P(A)}{P(B)} \quad (1)$$

Pada pengaplikasiannya rumus ini berubah menjadi:

$$P(Ci|D) = \frac{P(D|Ci) * P(Ci)}{P(D)} \quad (2)$$

Naïve Bayes Classifier adalah model penyederhanaan dari metode *Bayes* yang cocok untuk pengklasifikasian teks. Adapun rumusnya dipaparkan pada Persamaan (3) sampai (5):

$$VMAP = \underset{Vj \in V}{\arg\max} \frac{P(a1,a2,\ldots,an|vj)P(vj)}{P(a1,a2,\ldots,an)} \quad (3)$$

$$P(vj) = \frac{|doc_j|}{|Contoh|} \quad (4)$$

$$P(wk|vj) = \frac{nk+1}{n+|Konstanta|} \quad (5)$$

Dimana :

P(vj) : Probabilitas setiap dokumen terhadap sekumpulan dokumen
P(wk|vj): Probabilitas kemunculan kata wk pada suatu dokumen dengan kategori klas vj.
|docs| : Frekuensi dokumen pada setiap kategori
|Contoh| : Jumlah dokumen yang ada
Nk : Frekuensi kata ke-K setiap kategori.
Kosakata: Jumlah kata pada dokumen tes.

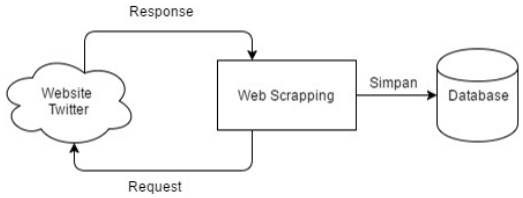
3. Metodologi

Dalam metode penelitian ini akan menjelaskan langkah-langkah yang dilakukan untuk merancang aplikasi sentiment analisis twitter sebagai berikut :

3.1. Metode Pengumpulan Data

Pada penelitian ini pengumpulan data dimulai dengan penarikan data *tweet* dari *website twitter* yang kemudian disimpan ke *database*. Penarikan data *tweet* dilakukan dengan menggunakan fasilitas *Web Scrapping*.

Web Scrapping merupakan suatu teknik untuk mengutip dan mengekstraksi data atau informasi dari suatu *website* dengan menggunakan *low-level HTML*. *Web Scrapping* ini mengambil data kotor secara *realtime* dari *website twitter,* yang selanjutnya akan dipilih menjadi data *tweet* bersih. Data *tweet* ini akan disimpan di dalam *database*. Berikut gambar proses pengumpulan data:



Gambar 1. Proses Pengambilan Data Twitter

Data yang diambil dari server twitter diperoleh dari *tweet* yang dibuat mulai november 2012 sampai januari 2017 dengan kata kunci pencarian samsat malang kota. Setelah proses pengambilan data selesai dilakukan. Data *tweet* akan digunakan dengan cara membagi seluruh data yang diambil menjadi 2 bagian, yaitu data

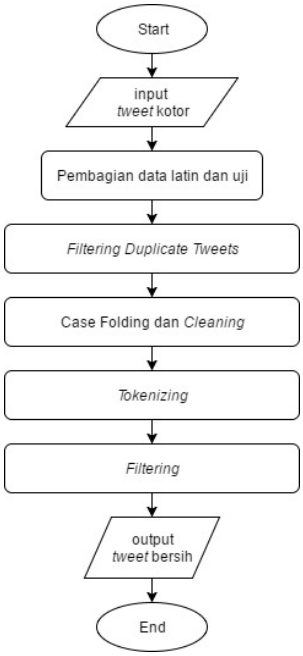
latih sebanyak 80 % dan data uji sebanyak 20 % Vinet (2011). Untuk proses pembuatan data latih, pengklasifikasian data dilakukan secara manual kedalam tiga kategori, yaitu positif, negatif, dan netral.

3.2. Metode Pengolahan Data

a. Preprosessing

Preprocessing ini dilakukan untuk menghindari data yang kurang sempurna, gangguan pada data, dan data-data yang tidak konsisten. Tahapan pada text preproseessing yang dilakukan adalah:

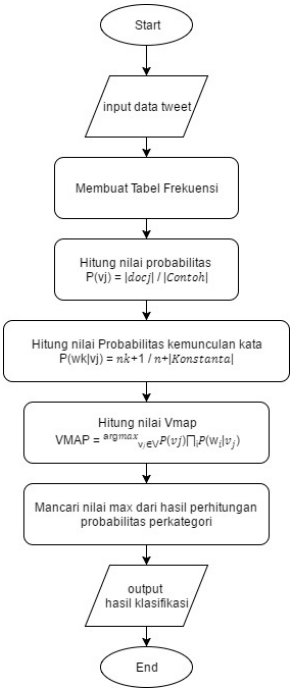
1. Melakukan *filtering duplicate tweets* adalah tahapan dimana *tweet* yang mempunyai isi sama akan dihapus untuk menghindari duplikat isi dari *tweet*.
2. *Case folding* dengan mengubah semua huruf dalam dokumen menjadi huruf kecil. Hanya huruf "a" sampai dengan "z" yang diterima.
   Contoh : Sb menambah wangi segar !!
   Menjadi : sb menambah wangi segar
3. *Cleaning* adalah tahap dimana karakter selain huruf dihilangkan dan dianggap *delimiter* dan menghapus juga *URL*, *mention* dan *hastag*.
   Contoh :@ aku https://path.com/p/3pB4Qs
   Menjadi : aku
4. Tahap *tokenizing / parsing* adalah tahap pemotongan string *input* berdasarkan tiap kata yang menyusunnya.
   Contoh : sb menambah wangi segar
   Menjadi : sb | menambah | wangi | segar
5. *Filtering* adalah tahap mengambil kata-kata penting dari hasil token. Bisa menggunakan algoritma *stoplist* (membuang kata yang kurang penting) atau wordlist (menyimpan kata penting). *Stoplist/stopword* adalah kata-kata yang tidak deskriptif yang dapat dibuang. Contoh stopwords adalah "yang", "dan", "di", "dari", "dengan" dan seterusnya Triawati (2009).
   Contoh : | sb | menambah | wangi | segar |
   Menjadi : | sb | wangi | segar | Senang



Gambar 2. Tahapan Preprocessing

b. Klasifikasi Model

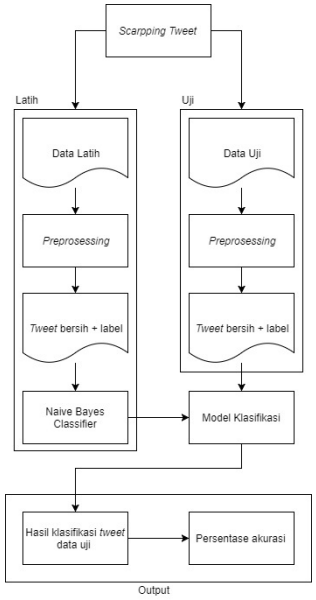
Perancangan *flowchart* ini bertujuan untuk memberi gambaran bagaimana proses klasifikasi *tweet* dengan perhitungan algoritma *Naïve Bayes Classifier*, mulai dari memasukkan data *tweet* yang akan diuji hingga aplikasi dapat menampilkan hasil akhir klasifikasi *tweet*. *Flowchart* dapat dilihat pada Gambar 3.



Gambar 3. Proses Naïve Bayes Classifier

3.3. Kerangka Konsep Sistem

Kerangka konsep penelitian ini dibuat untuk menggambarkan tahapan proses yang dilakukan aplikasi sentimen analisis mulai dari pengambilan data, preprosessing, sampai perhitungan klasifikasi dan akurasi. Kerangka konsep dapat dilihat pada Gambar 4.



Gambar 4. Kerangka Konsep Proses Pembuatan Sistem

Gambaran output sistem yang dibuat berupa jumlah kategori (positif, neagtif, netral) hasil dari klasifikasi yang dilakukan oleh sistem terhadap data *tweet* yang diuji dan penyajian akurasi kebenaran klasifikasi kategori dalam bentuk presentase.

3.4. Perhitungan Klasifikasi

Setelah dilakukan data preparation, selanjutnya melakukan proses text mining itu sendiri. Proses ini terbagi menjadi tahapan yaitu: pengolahan *tweet*, transformasi teks kedalam bentuk, dan perhitungan. Sistem akan melakukan pengolahan terhadap data masukan berupa data teks dari *tweet*.

Berikut contoh perhitungan yang diambil dari satu *tweet* inputan sebagai berikut: "Semoga diperlancar (????) (@ Samsat kota Malang) http://4sq.com/12wYt3d?"

Hasil pengolahan kata dapat dilihat pada table 1.

Tabel 1. Contoh Hasil Preprosessing *Tweet*

kategori	Teks	Jumlah Kata
Kotor	Semoga diperlancar (????) (@ Samsat kota Malang) http://4sq.com/12wYt3d?	8
Bersih	semoga, diperlancar	2

Tahap selanjutnya yaitu melakukan proses perhitungan menggunakan metode Naïve Bayes dengan menggunakan data *tweet* yang diambil dengan kata kunci samsat malang kota.

Data dibagi menjadi 2 bagian yaitu data training yang terdiri dari 8856 kata yang telah diketahui kategorinya masing masing terdiri dari 2866 kata negatif, 2094 kata positif, dan 3896 kata netral. Berupa 859 *tweet* dengan kategori positif 203 *tweet*, negatif 284 *tweet*, netral 370 *tweet*. Data uji merupakan sebuah tweet yang belum diketahui kategorinya.

Berikutnya akan diberikan contoh penggunaan algoritma *Naïve bayes* untuk klasifikasi sentimen. Misalkan untuk *tweet* berikut ["Semoga diperlancar (????) (@ Samsat kota Malang) http://4sq.com/12wYt3d?"], yang sudah melalu proses text mining sehingga mendapatkan hasil term frequency (TF).

Hasil perhitungan Naïve Bayes untuk *tweet* tersebut seperti terlihat pada Tabel 2.

Tabel 2. Contoh Frekuensi Kata Uji Pada Data Latih

no	kata	Training negatif	Training positif	Training netral
1	semoga	0	8	0
2	diperlancar	0	2	0

Dari Tabel 2 diketahui: Jumlah Term Frekuensi (TF) keseluruhan dari N_k di setiap kategori. Dan dari data latih didapatkan (n_c | Positif) = 2099, (n_c | Netral) = 3896), (n_c | Negatif) = 2866) dan Jumlah Kosa kata Training = 2023.

Dari nilai-nilai tersebut, dapat dicari nilai nilai probabilitas keyword dengan menggunakan rumus P(Xi|Vj) dan probabilitas kategori dokumen P(Vj) yaitu:

P(Vj | Positif) = 203/859 = 0.23632130384167638
P(Vj | Negatif) = 203/859 = 0.33061699650756693
P(Vj | Netral) = 203/859 = 0.4307334109429569

P(semoga | positif) = (8+1)/(2094+2023)
= 0.002183406113537118
P(semoga | negatif) = (0+1)/(2866+2023)
= 2.0454080589077522E-4
P(semoga | netral) = (0+1)/(3896+2023)
= 1.6894745734076703E-4

Hasil dari perhitungan P(Xi|Vj) kategori positif dapat dilihat pada Tabel 3.

Tabel 3. Contoh Hasi Perhitungan Probabilitas Kata

kata	N_k positif	P(Xi|Vj) positif
semoga	8	0.002183406113537118
diperlancar	2	7.27802037845706E-4

Berdasarkan nilai probabilitas pada tabel dapat dihitung nilai Vmap:
VMAP = $\arg max_{vj \in V} P(vj) \prod_i P(ai|vj)$ = (0.23632130384167638)*(0.0021834061135371 18)*(7.27802037845706E-4) = 3.755352107474374E-7

Hasil akhir perhitungan Naïve bayes untuk contoh *tweet* di atas dapat dilihat pada Tabel 4.

Tabel 4.Hasil Akhir Klasifikasi

Tweet	Semoga diperlancar (????) (@ Samsat kota Malang) http://4sq.com/12wYt3d?
Training Positif	3.755352107474374E-7
Training Negatif	1.3832003867221386E-8
Training Netral	1.2294528564035867E-8
Klasifikasi	positif

Perhitungan pada Tabel 4 memperlihatkan proses penentuan klasifikasi sentimen didasari oleh hasil perhitungan probabilitas dengan nilai tertinggi yaitu 3.755352107474374E-7, sehingga *tweet* diatas dikategorikan sebagai *tweet* "Positif".

Langkah-langkah perhitungan di atas kemudian diterapkan pada pengembangan aplikasi sentimen analisis sederhana berbasis desktop.

4. Pengujian dan Pembahasan

Pada bab pengujian dan pembahasan ini akan dilakukan tahapan untuk menguji hasil dari implementasi sistem yang telah dilakukan.

4.1. Pengujian Fungsional

Pengujian sistem ini dilakukan dengan cara menjalankan aplikasi secara detail pada setiap menu yang ada, dengan tujuan untuk mengetahui menu atau fitur mana yang sudah berfungsi dengan baik maupun yang tidak berfungsi sesuai dengan sebagaimana mestinya.

4.2. Pengujian Akurasi

Dari pengujian yang telah dilakukan dengan 2 jenis data latih didapatkan jumlah data *tweet* yang sesuai dengan klasifikasi manual, untuk pengujian 1 didapatkan klasifikasi benar sejumlah 156 *tweet*, klasifikasi salah sejumlah 59 *tweet*. Untuk pengujian 2 didapatkan klasifikasi benar sejumlah 161 *tweet*, salah 54 *tweet*.

4.3. Pembahasan

Dari hasil perhitungan *Precision*, *Recall*, dan *Accuracy* pada setiap kategori klasifikasi, nilai pada pengujian skenario 1 dan pengujian skenario 2 dari masing –masing kategori positif, negatif , dan netral dapat dilihat pada Tabel 5.

Tabel 5. Hasil Pengujian

K	Skenario Uji 1			Skenario Uji 2		
	P	R	A	P	R	A
Pos	0.45	0.43	0.81	0.50	0.32	0.82
Neg	0.53	0.56	0.89	0.66	0.64	0.92
Ne	0.85	0.86	0.80	0.83	0.91	0.80

Dari hasil pengujian diatas dapat diambil kesimpulan, bahwa hasil akurasi dari uji skenario 1 pada setiap kategori positif, negatif, netral masing-masing sejumlah 81%, 89 %, 80 %, dan hasil akurasi dari uji skenario 2 pada setiap kategori positif, negatif, netral masing-masing sejumlah 82%, 92%, 80%. Hasil akurasi tertinggi diperoleh dari uji skenario 2 dengan mempergunakan data latih yang memiliki jumlah data *tweet* yang sama pada setiap kategorinya. Hasil *precision* dan *recall* dari uji skenario 2 mendapatkan nilai yang lebih bagus dari uji skenario 1, dikarenakan data latih yang diguakan memiliki jumlah data yang berimbang disetiap kategorinya. Hasil dari uji skenario 2 didapatkan nilai *precesion* dan *recall* tertinggi sebesar 83% dan 91% pada kategori netral.

Dari hasil uji coba yang telah dilakukan diperoleh bahwa hasil implementasi algoritma naïve bayes classifier untuk klasifikasi data tweet pada aplikasi sentimen analisis ini sudah berjalan sesuai dengan hasil yang diinginkan.

5. Kesimpulan

Dari hasil penelitian terhadap analisis sentimen data twitter dengan menggunakan algoritma *naïve bayes classifier* terhadap Sistem Administrasi Manuggal Satu Atap (SAMSAT) Malang Kota dapat disimpulkan sebagai berikut :

1. Metoda *Naïve Bayes Classifier* dapat diterapkan sebagai metode untuk melakukan klasifikasi sentimen analisis.
2. Pengumpulan text berupa *tweet* dari twitter dapat mempergunakan Web Scrapping untuk alternatif.
3. Aplikasi Sentimen Analisis yang dikembangkan dianggap cukup memadai, dikarenakan hasil uji akurasi klasifikasi yang dilakukan oleh aplikasi pada data uji dan dilakukan dengan 2 variasi skenario, hasil akurasi dari uji skenario 1 pada setiap

kategori positif, negatif, netral masing-masing sejumlah 81%, 89%, 80%, dan hasil akurasi dari uji skenario 2 pada setiap kategori positif, negatif, netral masing-masing sejumlah 82%, 92%, 80%. Hasil akhir menghasilkan nilai akurasi tertinggi diperoleh skenario uji 2 yaitu pada setiap kategori positif, negatif, netral masing-masing akurasi didapat sebesar 82%, 92%, 80%, dengan jumlah kata latih 200 *tweet* negatif, 200 *tweet* positif, dan 200 *tweet* netral.

4. Semakin banyak data latih dan mempunyai jumlah data latih dengan jumlah kategori yang sama antar kategori maka nilai sama antar kategori maka nilai sentimen yang didapat akan semakin akurat.

6. Saran

Dari kesimpulan yang telah diuraikan maka terdapat saran yang perlu disampaikan untuk penelitan selanjutnya yaitu penelitian selanjutnya dapat menggunakan metode alternatif lain untuk mengklasifikasi teks dan fitur preprosessing yang lebih lengkap, serta menggunakan data latih yang lebih banyak sehingga hasil yang didapat semakin akurat.

Daftar Pustaka:

Hemalatha, I., Varma, P.G., dan Govardhan, A., (2012), *Preprocessing the Informal Text for Efficient Sentiment Analysis, International Journal of Emerging Trends & Technology in Computer Science (IJETTCS),* Vol. 1, July – August 2012, ISSN 2278-6856.

Josi, A. L.A. Abdillah, Suryayusra, (2014), *Penerapan Teknik Web Scraping Pada Mesin Pencari Artikel Ilmiah*.

Liu, B, (2012), *Opinion Mining*. Chicago, United States of America.

Puntoadi, Danis, (2011), *Menciptakan Penjualan Melalui Social Media*. Jakarta: PT Elex Komputindo

Triawati, C, (2009), *Text Mining*. Bandung, Jawa Barat, Indonesia

Turland, M, (2010), *Php| Architect's Guide To Web Scraping With PHP*. Introduction-Web Scraping Defined, str, 2.

Yadav, V, (2011), *How Big Should The Training Set Be In The Naive Bayes Text Classification?* dari https://www.quora.com/How-big-should-the-training-set-be-in-the-Naive-Bayes-text-classification. Diakses 03 Februari 2017.